

EFFECTS OF EXAMINEE'S NATIONALITY, CONTENT
CORRECTNESS, AND MECHANICS OF CONTENT
ON RATINGS GIVEN TO ESSAY EXAMINATIONS

BY

BEHROKH AHMADI

A DISSERTATION PRESENTED TO THE GRADUATE COUNCIL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL
FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1982

Copyright 1982

by

Behrokh Ahmadi

ACKNOWLEDGEMENTS

I am deeply indebted to Dr. Linda Crocker, chairperson of my doctoral committee, for her continued encouragement and support. I shall always be extremely grateful to her. Whatever skills I have developed as a researcher are in large part due to her advice and guidance.

I would also like to express my appreciation to the members of my committee, Dr. Robert Soar and Dr. Walter Cunningham, for their suggestions and editorial comments on this dissertation. Special thanks also are extended to Dr. John Bengston, Dr. James Algina, and Dr. Steve Olejnik for their assistance and suggestions. I would like to thank all the raters who participated in this study.

Finally, I would like to express my grateful appreciation to my parents without whose support I would not have been able to reach this point in my graduate studies. My special thanks to my sisters who lived through this tremendous experience with me, and thanks to my brothers for their encouragement. I extend my thanks to my friends for their support.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS.....	iii
ABSTRACT.....	vi
CHAPTER	
I INTRODUCTION.....	1
Problem Description.....	1
Purpose of This Study.....	3
Theoretical Rationale.....	5
Educational Significance.....	7
Summary.....	9
II REVIEW OF LITERATURE.....	10
Introduction.....	10
Essay Examination.....	10
The Writer.....	12
Structure of the Grading Situation.....	14
Scoring Guide.....	14
Context Effect.....	15
The Rater.....	16
Rater Characteristics.....	16
Characteristics of Essay Response.....	18
Rater's Perception of Examinees.....	22
Testing and Grading Foreign Students.....	24
Foreign Students in U.S. Educational Setting	24
Language and Communication Problems of	
Foreign Students.....	25
Summary.....	28
III METHODOLOGY.....	31
Instrumentation.....	31
Generating the Essay Item and Responses.....	31
Construction of Order Effect.....	33
Pilot Studies.....	35
Scoring Procedure.....	35
Subjects.....	35
Assumption about the Rating Procedure	
and Raters.....	37
Design of the Study.....	37
Split Plot Design.....	37
Summary.....	43

	Page
CHAPTER	
IV RESULTS.....	44
Introduction.....	44
Results of Four-way Factorial Analysis from a Split Plot Design.....	46
Choice of Error Term for Simple Main Effect....	50
Conservative Degrees of Freedom.....	51
Results of Split Plot Design Analysis for Different Content Accuracy Levels.....	53
Summary.....	69
V DISCUSSION.....	71
Mechanics.....	72
Nationality.....	73
Conclusion.....	77
Limitation of Study.....	79
Suggestions for Future Research.....	80
Summary.....	81
APPENDIX	
I SAMPLE OF ESSAY ITEM USED IN THE STUDY.....	84
II SAMPLE OF SOME ESSAY ITEM RESPONSES USED IN THE STUDY.....	86
III SAMPLE OF SCORING INSTRUCTION.....	90
IV SAMPLE OF SCORING GUIDE USED FOR RATING OF ESSAY ITEM RESPONSE IN THE STUDY.....	92
V SAMPLE OF THE LETTER OF INVITATION FOR PARTICIPATION IN THE STUDY.....	94
REFERENCES.....	95
BIOGRAPHICAL SKETCH.....	99

Abstract of Dissertation Presented to the Graduate Council
of the University of Florida in Partial Fulfillment of
the Requirements for the Degree of Doctor of Philosophy

EFFECTS OF EXAMINEES' NATIONALITY, CONTENT
CORRECTNESS, AND MECHANICS OF CONTENT
ON RATINGS GIVEN TO ESSAY EXAMINATIONS

By

Behrokh Ahmadi

May 1982

Chairperson: Linda M. Crocker
Major Department: Foundations of Education

Reliability and validity of essay examinations can be seriously affected by the subjectivity of the scoring procedure. The ratings of essay responses are affected by variables which are unrelated to accuracy of the essay's content. The problem of obtaining reliable achievement measures from essay examinations is especially difficult when the examinees are not native English speakers.

A review of literature revealed numerous studies conducted to examine the effects of handwriting, grammatical errors, appearance of the essay, raters' expectations of examinees' achievements, raters' background, and context in which the essays are read on raters' judgments of essay responses, but there had been no study on the effect of knowledge of examinees' nationalities on raters' judgments of essay responses.

An experimental study was designed to test the effect of raters' knowledge of examinees' nationalities, content correctness of the responses, mechanical accuracy, and order of presentation of essay on ratings of essay responses. Sixty raters graded 48 essay responses to an essay item in educational psychology. These 48 essay responses were divided into four sets, 12 essays in each set. Each rater rated 12 essays which had all possible combinations of the four variables.

Data from a four-way factorial split plot design were analyzed. A significant four-way interaction was found between content, mechanics, nationality, and order. Separate three-way factorial analyses were then conducted for each level of content accuracy; significant three-way interactions were followed by separate analysis for each level of mechanical accuracy.

The effect of mechanics was significant at only one level of order, and the effects of nationality were significant at all four levels of order. When the essays were partially correct or incorrect in the content and had many mechanical errors, the effects of nationality were significant at first and third levels of order. When the essays were partially correct or incorrect in content and had few mechanical errors, the effects of nationality were significant at second and fourth levels of order. However, the higher mean was always assigned to the nationality which appeared first in the order of presentation. Implications for scoring essay examinations of foreign students were discussed.

CHAPTER I

INTRODUCTION

Problem Description

Essay examinations have been criticized during the past fifty years by a number of researchers such as Huck and Bounds, who cite the problems of scoring such tests objectively. They have pointed out that "whenever an essay test is used, factors which are extraneous to purpose(s) of the test may enter into the grading procedures" (1972, p. 283). Marshall (1967) has suggested that the influence of such extraneous factors reduces the meaning of the grades given and foils the purpose for which they are assigned. He has stated that

The general purpose of giving grades is to provide a description of an individual's achievement in a given area. For an essay examination to be useful for measuring achievement in an academic course, it should be scored with reference to specific objectives which are germane to the subject matter being measured. (p. 375)

He suggests that raters of essay tests should grade the content of the essay only on the basis of achievement in the subject matter and all the other factors should be ignored. Yet most raters seem unable to do this (Scannel and Marshall, 1966).

A number of studies have shown that ratings of essay responses are affected by three sets of variables unrelated to the accuracy of the response: (1) the characteristics of the written piece or appearance of the essay examination, which includes variables such as handwriting, spelling, grammatical accuracy and compositional fluency (Bean, 1953; Marshall, 1967; Chase, 1968; McColly, 1970; and Harris, 1977). (2) the order of context in which the essay responses are read (Hales and Tokar, 1975; Hughes et al., 1980), and (3) personal characteristics of the examinee, which includes variables such as race (Wen, 1979). Wen found some racial bias when raters graded essays written by white, black and oriental students. Chase (1979) has even reported that expectations of students' performance can influence raters' essay grading. There is also evidence that the essay raters' knowledge of examinee characteristics, for example, physical attractiveness, will affect their expectations of examinee performance (Clifford and Walster, 1973). With all these factors having the potential to influence essay grading, the virtue of measuring educational achievement by using an essay examination is dependent on the quality of the grading process.

In view of the finding of Wen (1979) and Chase (1979) one obvious characteristic that may affect raters' judgments of an essay response is examinee nationality. It seems likely that teachers' expectations of foreign students'

performance differ from their expectations of American students'. How much (if any) this difference in expectations influences their judgments of essay responses has never been studied experimentally.

Purpose of This Study

The purpose of this study was to determine how some specific paper and examinee characteristics, unrelated to content accuracy, may interact to affect ratings given to essay item responses in educational psychology. The variables of interest in this study were the nationality of examinee, level of mechanical errors in the essay response, content correctness, and serial position of papers containing these factors in combination.

The following research questions were investigated:

1. Will knowledge of examinee's nationality affect raters' judgments of essay item responses?
2. Will level of mechanical accuracy in the essay responses affect raters' judgments of essay responses?
3. Will level of content accuracy affect raters' judgments of essay responses?
4. Will the order of presentation of different levels of the nationality and mechanic factors affect raters' judgments of essay responses?

5. Will any of the following two-way interactions affect the raters' judgments of essay responses:
- a. interactive effect of nationality and mechanical accuracy effects?
 - b. interactive effect of nationality and content correctness effects?
 - c. interactive effect of mechanical and content correctness?
 - d. interactive effect of nationality and order effects?
 - e. interactive effect of mechanical accuracy and order effects?
 - f. interactive effect of content accuracy and order effects?
6. Will any of the following three-way interactions affect the raters' judgments of essay responses?
- a. interactive effect of nationality, mechanical accuracy and content accuracy?
 - b. interactive effect of nationality, mechanical accuracy, and order of presentation?
 - c. interactive effect of mechanical accuracy, content accuracy, and order of presentation?
 - d. interactive effect of nationality, content accuracy, and order of presentation?
7. Will the interactive effect of nationality, mechanical accuracy, and content correctness, and order affect the raters' judgments of essay responses?

Theoretical Rationale

Essay tests are known to have two shortcomings, namely validity and reliability. Reliability is considered to be the more basic issue of the two. According to McColly (1970), "If a test is totally non-valid, but is nonetheless reliable in all respects, it may be worse than useless because of the likelihood of its misuse" (p. 149). Hogan and Mishler believe that "one of the greatest fears that researchers have is the judgments of the quality of students' writing are capricious, unstable, or to use the psychometric term, unreliable. . . . without a certain amount of agreement in our appraisals, essay scores don't mean much of anything" (1979, p. 143).

Even though the investigations to identify the sources of unreliability of essay examinations have been ongoing for some time, not all the sources have been studied. Braddock, Lloyd-Jones, and Schoer (1963) have recognized four sources of variation, which would be accounted for when rating essay exams. These sources are the writer variable, the assignment variable, the rater variable, and the colleague variable. They have described the writer variable as any variable which will cause the student not to write at his capacity level. The assignment variable has four aspects: the topic, the mode of discourse, the time afforded for writing, and the examination situation; it is any outside variable which may affect the examination process. The rater variable has been defined as the "tendency of a rater

to vary in his own standards of evaluation" (p. 10). The fourth variable is the colleague variable, which has been defined as "the tendency of several raters to vary from each other in their evaluations" (p. 11).

The focus of this study is on rater-variation. The main source of undesirable rater variability is rater bias, which may be caused by the rater's characteristics, the rater's beliefs about the characteristics of the writer or the rater's reaction to characteristics of the essay response. Bias caused by raters' beliefs about examinees can be one of the most important sources of the unreliability in essay grading. Ebel (1972) points out the freedom that an examinee has in an essay examination to express his individuality in his answer will influence the judgment of the rater. He says that

The scorer's freedom to express himself in his evaluation also contributes to unreliability. This freedom also allows him to respond, in his evaluations, to what he already knows or believes about the student. A good answer from a poor student tends to be discounted. A poor answer from a good student tends to be evaluated more highly than its merits deserve. (p. 131)

Thus, raters' beliefs about past performance, educational background, or physical examinee characteristics, or social characteristics like ethnic background, race and original nationality of student, may affect the raters' judgments of an essay item response. This study was conceived to test the effect of this particular type of rater bias and its

interaction with another well-known source of rater bias in essay grading, mechanical errors. In addition the effect of content accuracy of the responses and order of presentation of different combinations of writer and paper characteristics were studied.

Educational Significance

In the past two decades enrollment of international students in colleges and universities in the United States has steadily increased. Many of the difficulties these students experience during the first two years are related to their ineffective communications in the English language. Robertson and Rogers have noted that many foreign students have difficulty during their first academic year in the United States; these students have indicated that "inadequate preparation in English contributed substantially not only to their lack of academic success but also to problems they encountered in acculturation" (1979, p. 57).

International students need time to adapt to the new culture and to develop an oral language base in a new language. Troyanorich has indicated that most international students feel "neither the psychological nor the practical need to learn to write the foreign language" (1974, p. 435) (emphasis added). Lack of ability to write fluently does not mean that the international students are unable to learn or achieve. Ability to write an essay in a second language can develop only after the student has developed an adequate

vocabulary and mastered grammatical structure of the second language. For this reason international students are particularly ill-equipped to display their mastery of subject matter in essay testing situations. Because most instructors recognize this, it is likely to have some impact on their ratings when grading the essay responses of international students. The impact may be caused either by the instructor's expectation or by the appearance of the essay response. One possible instructor's expectation of foreign students may be that they should be able to write as well as native speakers. In contrast the expectation may be that foreign students cannot write well and cannot express their thoughts efficiently. An instructor may also be affected by the appearance of the essay response; in this case the instructor may penalize the content of the response for bad writing or in contrast may give more credit to content than should be given. In either case the foreign students have a handicap in writing in a second language which may appear in their essay responses. Thus, there is need to study the effects of rater reactions to students' nationality along with other possible sources of unreliability of the essay examination scores. Teachers' knowledge of students' nationality may be another source of unreliability in the evaluation process. This information would be valuable for future efforts to improve evaluation of students' achievement in bilingual settings.

Summary

Essay examination responses have characteristics, unrelated to content accuracy, which are apparent to raters and have been shown to affect their judging behavior (Scannel and Marshall, 1966).

A number of studies have shown that ratings of essay responses are affected by three sets of variables unrelated to the accuracy of the responses: (1) the characteristics of the written piece or appearance of the essay examination, which includes variables such as handwriting, spelling, grammatical accuracy and compositional fluency (Bean, 1953; Marshall, 1967; Chase, 1968; McColly, 1970; and Harris, 1977); (2) the order of context in which the essay responses are read (Hales and Tokar, 1975; Hughes et al., 1980); and (3) personal characteristics of the examinee, which include variables such as race (Wen, 1979).

In the present study the effects of some paper and examinee characteristics on the ratings of essay test responses were examined. Specifically, the effects of content correctness, level of mechanical accuracy, knowledge of examinees' nationality and the presentation order were investigated, with attention to their possible interactions.

CHAPTER II

REVIEW OF LITERATURE

Introduction

The review of the literature has been divided into two main sections: the essay examination, and testing and grading foreign students. There are sub-sections under each of these sections.

Essay Examinations

Reading and writing are tools of passing on the knowledge of sciences and arts and of grasping the knowledge of the world. Wilkinson emphasizes writing by saying that

. . . it is difficult to be sure we have thought it clearly until we have written it and communicated it successfully to others. Writing clarifies thought, firms it up and makes it a more suitable foundation on which to build subsequent thought. (1974, p. 25)

One use of the written communication process is the essay examination in which the writer presents the level of achievement in a subject matter. Essay examinations for formal assessment of achievement have been used since at least 2200 B.C. when civil servants of the Chinese dynasty were hired on the basis of a grueling essay

testing (Dubois, 1964). Although the popularity of this examination form waned somewhat with the advent of objective examinations, measurement experts still recognize the value of this type of examination. McColly (1970) has presented uses of writing exams as "measures of writing ability, measurement tools for thinking ability, or ability to put subject matter mastery into the mode of written discourse" (p. 149). In an essay examination the student is permitted to create a response, and answers given to a question differ according to different writing abilities as well as to different levels of knowledge about the subject matter. One major reason for using essay examinations is the affirmed power of the essay tests to expose the student's ability to organize, integrate, and synthesize his knowledge (Yeasmeen and Baker, 1973 and Ebel, 1972). Bean (1953) also has noted that the essay exam, due to the amount of writing required, discouraged students from cheating, and encouraged better study habits.

Wong (1973) has suggested that another major reason that essay exams are used is one of their least educationally prominent properties; there is a widespread belief that they are easy to construct. For teachers with limited time this property usually overcomes the reservations about other weaknesses of this type of examination.

There are some notable weaknesses of essay examinations that make these tests less than ideal. The most common is the unreliability of the grades assigned by raters. According to Hogan and Mishler (1979), the judgment of essay responses is an unstable process and without a reasonable agreement in raters' assessment, the essay scores are meaningless. The validity is also a consideration. Use of a test with an average reliability which is totally non-valid might be worse than using a test with no reliability (McColly, 1970).

As cited in Chapter I the sources of unreliability and invalidity in essay examination scores are those sources of variation unrelated to examinee's knowledge in the content area. Braddock, Lloyd-Jones, and Schoer (1963) have suggested a system for categorizing these sources of variation: 1) the writer, 2) the rater, 3) the assignment, and (4) colleague variable. A modified version of this set of categories will be used for the following review of literature on factors affecting the essay examination. The set of categories will include 1) the writer, 2) the structure in grading situation, and 3) the rater. There will be sub-categories under each category.

The Writer

The issue of bluffing has been mentioned by Ebel (1972) as one weakness of essay examinations; a student caught with a question that he is unable to answer can

direct his answer to a similar question which he knows more about and can answer. If this is done well, and a complete answer is given, the reader may miss the change and grade the response as being a relevant answer. The tendency to bluff probably varies widely among examinees and thus may be an undesired source of variation in grading essay responses. No empirical studies were found, however, to support this reasonable contention.

Motivation is an important factor which may vary among writers. What, if anything, has motivated the student to write at his best? There is no way to know if the student is using his full potential in writing situations (McColly, 1970). Another possible factor is the writer's reaction to distractions in the testing situation. Some writers are distracted by any outside element, and this may effect the writing process. Again, no empirical research of these factors was found but McColly has noted that if a student is distracted by some outside factor "who is to say that the measure of his performance should not reflect such distractibility? After all, distractions are very much a part of life, especially student life" (1970, p. 149).

Zimmerman studied the effect of eleven black college freshmen's self concepts of writing ability on their writing achievement. He found that "students who had high self assessments tended to write more confidently and more systematically and spend more time revising than did those with low self assessment" (1977, p. 7132-A). This

would indicate that, for some students at least, personality characteristics may influence their performance on essay tests.

Structure in the Grading Situation

Scoring Guide

Using a scoring guide is known to be one factor which can improve the reliability of essay exams. Chase (1968) investigated the effects of quality of handwriting, spelling accuracy and use of a scoring key on scores given to essay items. He asked 128 graduate students in classes in educational measurement to score the responses to two items. He found that having a scoring key increases the grade given to an essay and that with a scoring key the raters could keep a better track of the points to be graded. Wong studied the effect of use of a scoring guide on the reliability improvement. Fifty-two raters judged 10 essays under two conditions, using a scoring guide and not using a scoring guide. The result of the analysis indicated that a "scoring guide substantially improved the reliability of content grading process" (1973, p. 127).

Hogan and Mishler have claimed that "we've studied the problem of reliability extensively and have found amazing agreement on the quality of writing if there is consensus among teachers about what to look for in an essay" (1979, p. 143). They report no matter if teachers were college faculty, high school teachers or teacher trainees the guidelines helped the objectivity of essay scoring.

Context Effect

Another source of error in essay grading recognized by McColly (1970) is the order in which the essays are read. Context position of an essay exam seems to have an affect on the grade assigned to it. Hales and Tokar (1975) studied the effect of the quality of the paper preceding an essay response on the grade assigned to the response. Twenty-six responses to a question on the American Revolutionary War appropriate to fifth and sixth graders were graded by a panel of six doctoral candidates in elementary education. Hales and Tokar have reported that "the data suggested that a block of very good responses depresses the grades assigned to subsequent responses. Conversely, a block of poor paper enhances the grades assigned subsequent papers" (p. 116). Hughes, Keeling and Tuck (1980) have also evaluated the problem of context position in essay grading. In their study, 25 experimental teachers scored essays written by 38 high school students aged 13-14 years. They have pointed out that findings indicate that "context effects in grading are persistant and need to be guarded against even after a number of essays of varying quality have been read and irrespective of whether analytic or holistic scoring procedures are used" (p. 132).

The Rater

Research has long indicated that raters are affected by some unrelated factors when they judge the content of an essay examination. Readers of essay examinations are distracted from the accuracy of the responses by factors related to the writer or the factors related to the written piece. Furthermore, when different raters read the exam, personal characteristics of the raters also contribute to score variance. Research findings in these areas are developed in this section.

Rater Characteristics

Raters' competence has been suggested as one possible source of unreliability of essay examination grades. Themier (1967) investigated differences in grading practices among teachers with differing levels of teaching experience and probationary-tenure status. His findings indicate that "differences in grading practices cannot be attributed to the probationary-tenure status of teachers, nor the number of years of experience in teaching" (p. 2922A).

Marshall (1967) found that the teachers with more experience were affected more by the errors of essay exams and that teachers who have training in the content area graded more severely than teachers who did not have training in content area. By contrast, Klein and Hart (1968) found that "professors in law assigned the same grades to essay responses (subject matter of essay was in law) as the

raters who did not have any law training. It appeared, therefore, that the raters were at least in part giving good grades to those papers that presented a persuasive 'common sense' answer to the question" (p. 205). McColly (1970) reported findings of research done at the University of Wisconsin indicating that interrater reliability was higher among a group of raters who had higher educational background than of the other raters with lower educational background. McColly also emphasized the effect of training by saying that ". . . it is plain that readers must be given the proper training and orientation, regardless of how knowledgeable they are . . . practice is indispensable in helping readers work up to the proper speed and rate" (p. 150).

The rater's mood at the time of grading has been noted as one factor affecting the essay evaluation process. Bean has stated that

The general feeling of well-being or the opposite mood in the grader is a normal human phenomenon that is difficult to deal with in essay evaluation. . . . The halo effect is closely related to this problem. When a teacher is in a pleasant mood while grading a paper, he is likely to find a good answer to one question and carry this impression over to the next question. (1953, p. 114).

According to Bean a number of studies have been made of judging one paper by several raters which show wide variation in the ratings, even when a scale which is more or less uniform has been used by raters. It happens that

sometimes one rater is consistently on the lenient side while another rater tends to be more exacting than the average. But more often, "there is a noticeable lack of any uniformity of pattern." This might be related to other personality characteristics of the rater. Snibbe (1970) studied the relationship between 60 professors' personalities and their actual grading behaviors (grade distribution data) for the courses that they taught during two years. She has reported that there is a significant relationship between professors' attitudes toward grading and their grading behaviors: students received significantly higher mean grades in courses taught by professors with liberal attitudes toward grading than students in courses taught by professors with conservative attitudes towards grading.

Characteristics of Essay Response

The characteristics of the written piece may also affect the raters' judgments. The length of response has been mentioned by Bean (1953). He has noted that "many of the factors that contribute to the subjectivity of rating are unconscious. The length of the answer is one that is often found to operate below the conscious level . . . it is a difficult fault to correct" (p. 114).

McColly (1970) pointed out that the most sources of unreliability and invalidity of essay exams are related to their appearance. He has indicated that "many readers

for many years have intuitively been certain of the bias which occurs from a paper's appearance, chiefly its handwriting" (p. 153). Most researchers who have investigated the effects of non-content related factors on essay scoring have looked at the handwriting variable and some researchers have investigated its interaction with other factors in causing bias in scoring of essay exams. Chase (1968) found that raters graded the poor handwriting papers very similar to good handwriting papers on the first item. "Poor handwriting papers got lower scores on the second item than on the first, while good handwriting papers got higher scores on the second item than on the first" (p. 318). This finding indicates an interaction effect due to the handwriting and the order in which the papers are read. Chase also found a significant effect due to the handwriting in this research. The graders were more generous in grading an essay with good handwriting. Manganel (1968) studied the effect of reading ease on grades given to essays. One hundred-twelve expository writing essays were graded by eleven teachers. The results of this experimental study indicated that "reading ease was a factor in essay grading process even when it is not part of grading criteria" (p. 836A).

McColly (1970) has commented on the effect of handwriting on validity and reliability of the papers. There is agreement among the raters about their reaction to the essay responses' appearance. He calls this situation

"ironic". When papers are read in handwriting, an inter-rater reliability will be high "because of the correlation of the appearance factor in the data. At the same time, the validity is actually lowered because handwriting ability and writing ability are not the same thing" (p. 153).

Huck and Bounds (1972) conducted an experimental study by asking 34 readers each to read two essays and assign grades to the essays. They studied the interaction between the writing clarity of readers and the neatness of essay papers which they read. They found that an interaction exists between the readers' handwriting neatness and reading ease of essay response being graded. Readers who had neat handwriting were "biased in the direction of giving lower grades to essays that contain messy handwriting. Readers who have messy handwriting, however, are not influenced by the neatness of essays" (p. 282).

The list of paper characteristics affecting the judgment of essay exams does not end with handwriting. In 1966 Scannel and Marshall studied the grading of essay examinations containing composition errors. The results of this study indicated that teachers are affected by the quality of composition in an essay examination. Scannel and Marshall have pointed out that "even when subjects are given specific directions to grade the paper only on the basis of content and are supplied with an outline of the desired content, they tend to assign lower grades to papers containing composition flaws than to those free from

such errors" (p. 128). In the following year (1967) Marshall used the data from the 1966 study and studied the effects of spelling, punctuation, and grammar on rating of essay exams; he found that papers with spelling errors were assigned the lowest mean grades, and papers with punctuation errors or combinations of errors did not differ significantly from mean grades assigned to the papers with no errors. In 1969, Marshall and Powers, using the results of two previous studies, decided to look at a new direction; interaction of handwriting neatness with only the extreme composition forms. Their somewhat surprising finding was "The rank of the writing of means from highest to lowest was: Neat handwriting, poor handwriting essay, typewritten essay and fair handwriting essay" (p. 100).

Similar studies have been conducted for grading of English compositions. Starring in 1952 chose 150 themes, which ranged in scores of 98-78, out of 2100 themes that were written in a written and spoken English comprehensive examination. He aimed to find the changes in ratings when some compositional elements were weakened. He re-wrote the essays to weaken the quality of grammar, sentence structure, diction, organization and content. He found that raters detect the grammatical errors easier than any other category. Grammatical errors were scored more severely than any other category. Results of the analysis did indicate that "the raters as a group did not make a distinction between weakness in sentence structure and

weakness in conventions of grammar" (p. 850). They also were unable to make a distinction between the weakness of content and weakness of organization. Spelling weakness was distinguishable for raters and they rated the papers with this weakness lower. The results of this investigation indicate that raters are not in agreement in making the distinction between the content of an essay or the other technical and mechanical characteristics of the composition. This disagreement contributes inevitably to error variance in scores of essay examinations.

Freedman (1979) studied what within the papers influences the judges. Ninety-six papers were judged by twelve readers. In her final findings, she has suggested that

The main results of the rewriting showed that the most significant influence on raters' scores was the strength of the content of the essay. The second most important influence proved to be the strength of the organization of the content. The third significant influence was the strength of the mechanics. Furthermore, the strength of the mechanics was most important when the organization was strong, and because the sentence structure alone was insignificant, the strength of the sentence was important only when the organization was strong. (p. 333)

Rater's Perception of Examinees

In addition to undesired variability due to raters' biases about paper characteristics, raters also seem to be affected by examinee's characteristics. Chase (1979) performed an experimental study to investigate the impact

of achievement expectations and handwriting quality on scoring essay tests. Sixty-two readers (graduate students) read responses to two essay items. He has reported that the expectancy has considerable influence on raters' judgment and its impact changed the "typical" relationship between handwriting quality and scores given to essay responses: "when the writing was less legible, the reader appeared to have depended more heavily on expectancy, with high expectancy group getting higher scores, and low expectancy group getting lower scores" (p. 141).

McColly (1970) has noted that the presentation of names on an essay answer is a source of scoring bias. He has pointed out that the effect of names, as a source of error, is a systematic factor and names should not appear on the essay response.

Wen (1979) investigated the existence of racial bias in teachers' rating of pupils writings. Three different writing themes were used in the study. One neutral theme, explaining an event, two themes about students' opinion of teachers, one with positive view and the other with negative view. Themes were accompanied by pictures showing the writer to be white, black or oriental. One hundred ninety-eight black raters read the themes. The findings included that

- 1) on the negative writing the black raters manifested strong racial bias in favor of black students and against white and oriental students,
- 2) on the neutral writing there was little evidence to support the existence of racial bias on the part of black raters, and 3) on the positive writing black readers' bias was significantly sex-directed rather than race-directed. (p. 17)

Testing and Grading Foreign Students

Foreign Students in U.S. Educational Setting

In most countries of the world today second language students comprise a rather large proportion of those receiving instruction in colleges or universities. Second language students are defined as students for whom the medium of instruction is not their native language. In most English-speaking countries they are foreign nationals, whose pre-university education has been under a different educational system with a different language. In many developing countries, at the higher level of education the language of instruction is an external one for the students, the native language of none of them (Diener and Kerr, 1979).

The number of foreign students studying in the colleges and universities of the United States has increased steadily in the past ten years. Diener and Kerr (1979) have pointed out that due to the vague definition of who is a foreign student, there is uncertainty about the number of foreign students in this country, but they have said that "foreign students were first reported to

be enrolled in U.S. community colleges in any sizeable number in 1968. Ten years later, estimates indicate a range of between 58,000 and 68,000" (p. 50). The last figure that they cite is 203,000 in 1976-1977. Florida institutions of higher education are attractive to foreign students. For example, Diener and Kerr further report that "Miami-Dade Community College in Florida leads the entire post-secondary education community in the enrollment of students from other lands" (p. 50). A substantial proportion of international students in the U.S. major in education. For example, according to Renner (1981) there are 191 foreign students in the college of education at Michigan State University out of a total of 1375 on the campus.

Language and Communication Problem of Foreign Students

Enrollment of foreign students in any American educational setting creates some problems and concerns. The most important is the English proficiency (or the lack thereof). Adequate knowledge of the English language is so fundamental for learning that universities and colleges are forced to stress a required level of English proficiency. Ure (1979) indicates that

...if we do not assume that most second-language students admitted will be able to use English reasonably well in working for their professional qualifications, we are forced into the position of saying, either that the admission policy is mistaken, or that national policy on the medium of instruction is misguided or misapplied. (p. 230)

Hosley and Meredith (1979) have noted that "...one admission requirement for all students whose native language is not English, is to demonstrate English proficiency...for many of these students the test of English as a foreign language (TOEFL) is the instrument chosen" (p. 209). They explain that "TOEFL is a standardized test designed to measure a person's proficiency in English, and consists of five subtests: listening comprehension, and writing ability" (p. 210). And they furthermore explain that "...the writing ability subtest contains two parts...this subtest is essentially tied to basic grammar and has format similar to other subtests" (p. 211). Chase (1972) has questioned whether recognition of an inconsistency in grammatical form tells us how the student can manage the form in his own writing. Thus it would appear that even though the foreign student may meet minimum admission requirements on the TOEFL, an objective test, this is no guarantee of ability to write fluently in the second language.

Payind (1979) has investigated the nature and extent of the academic, personal and social problems of Afghan and Iranian students in the United States. He has reported that

the most severe academic problems reported by both Afghan and Iranian student groups were: completing written examinations in the same length of time as American students do; improving English to the level necessary to pursue academic work; communicating thoughts in English,

presenting oral reports; competing with Americans for high grades; taking notes; and writing reports [emphasis added]. Academic problems seemed to be largely related to lack of proficiency with the English language and, to a certain extent, to the existence of differences between the educational systems of home countries and those of the United States. (p. 6)

Gipps and Ewen (1974) have indicated that concern about scoring essays written in a second language, saying "two of the most important considerations are the intelligibility and the structural complexity of the writing" (p. 121). The foreign students have a weakness in the second factor, and if the rater cannot detect the first factor without the interference of the second one, then the foreign students are being penalized for their language difference.

Gipps and Ewen (1974) direct the teachers who use essay examinations to use a scoring guide that not only includes the specific aspects of the writing but also an intelligibility factor. They point out that judgment of a written composition can be based on combination of factors such as length, accuracy, variety of content, linguistic maturity or literary style. They pointed out that relative weight should be given to each of these factors according to the aim of the essay examination. They have also pointed out that

. . . the effectiveness of written communication in a sense, depends largely on its intelligibility, involving a minimal degree of grammatical accuracy. An effective communication could also be said to be one where optimum use is made of the structural features of a language, i.e. (phrases, clauses, and sentences), to maximize the information expressed in a given number of words. (p. 121)

However, before generating guidelines for instructors to use in scoring essay exams of second-language students, it would be appropriate to know more about how instructors in higher education are currently scoring these students, and how their ratings are affected by grammatical flaws, when they know the examinee is writing in a second language. The present study addresses this question.

Summary

The use of essay examination as a formal type of achievement assessment goes back to 2200 B.C. One of the most noticeable weakness of essay examinations has been recognized to be the unreliability of scores given to such examinations. Sources of unreliability have been related to three factors: one is the writer; factors such as bluffing, motivation, and self-concept have been recognized to be related to students' performance in an essay examination. A second source of unreliability recognized is the structure in grading situations; use of a scoring key has shown to be an important factor in improving the

reliability of scores (Chase, 1968). The context position of an essay has been a significantly effective factor on ratings given to an essay response (Hales and Tokar, 1975). A third source of unreliability identified is the rater variable. Results of studies on the effect of raters' experience in the subject area have been contradictory (McColly, 1970; Themier, 1967). Rater's personality characteristics have been shown to affect their judgments of students' performance (Snibbe, 1970).

Some of the characteristics of essay responses such as handwriting, mechanical characteristics, spelling, and organization have been found to affect ratings given to essays (Scannel and Marshall, 1966; Freedman, 1979). The context in which the essays are read has been found to affect the ratings given to essays (Hales and Tokar, 1975; Hughes, Keeling and Tuck, 1980). Chase (1968) has found that order in which the essays are read interacts with handwriting. Another factor which can affect the rater's judgment of an essay response is the rater's expectation of the student's performance (Chase, 1979). Students' nationality is one variable which may affect teachers' expectations and therefore their ratings of essay responses.

One subpopulation of examinees most likely to have difficulty with essay examinations is the non-native English speaker; the number of these students involved in American colleges and universities has increased steadily in the last ten years (Diener and Kerr, 1979). The TOEFL,

used as a screening test for admission, for international students, does not test the student's ability to write in English (Chase, 1972). Thus, foreign students have identified their inadequacies in written communication to be severe problems in their academic adjustment (Payind, 1979). Gipps and Ewen (1974) have suggested that instructors may need to use special guidelines for grading written work of foreign students but offer no empirical basis for their recommendations. In view of the difficulties of writing in a second language (Payind, 1979) and the evidence that raters are affected both by the examinee writing and their expectations of examinee performance, the knowledge of students' nationality can be a serious source of bias in essay grading, which should be investigated.

CHAPTER III

METHODOLOGY

This study was designed to assess the effect of content accuracy, level of mechanical errors and knowledge of examinee nationality on ratings of essay examination responses.

Instrumentation

Generating the Essay Item and Responses

To develop the essay responses for this study, a number of essay items written for an introductory educational psychology course were reviewed. One item, which was judged to be well written and dealt with principles that are typically taught in introductory educational psychology, was chosen for this study (sample in Appendix I). The item dealt with behavior modification. After reviewing a number of actual student responses to this item, twelve fictitious responses were generated. Four responses were written at each of three levels of content correctness (highly correct, partially correct, and incorrect). The criterion to determine the levels of content correctness was the scoring guide given to raters (Appendix IV); the highly correct response had 5-6 of the steps described in the scoring guide, the partially correct response had 3-4 of these steps, and the incorrect

response had 0-2 of these steps. These twelve essay responses all dealt with the general problem of how to shape the smoking behavior of a roommate, but differed in their particular content. Papers were different in length; they used different subjects to shape the smoking behavior, like sister, husband, etc. and the responses described different reinforcers for the process of shaping.

Next the mechanical accuracy of written expression was modified. The mechanical accuracy was defined in terms of these categories: spelling, punctuation, grammatical accuracy and compositional fluency. Papers with few mechanical errors had 1-2 errors per hundred words in at least one of the categories (maximum of 7 errors per hundred words); papers with many mechanical errors had 1-3 errors per hundred words in each category (minimum of 11 errors per hundred words). In counting the number of errors, repeated misspellings of the same word were counted as only one error. Each of the twelve essays was written twice, once with many mechanical errors and once with few mechanical errors, resulting in a total of 24 essays.

Each of these 24 themes was then coded with an obviously "American" name and a duplicate paper was coded with a name of obviously foreign extraction. In choosing common American names, six American last names were selected from a telephone directory. The selection was made cautiously so that the

names would not represent any ethnic group in the United States (racial or geographical). For choosing foreign names, three geographical areas were chosen to represent the foreign nations: South American countries, Middle Eastern countries and Far Eastern countries. Students from different countries in the three areas were surveyed to collect some popular last names. Finally, six names were randomly selected from the list of names given (two from each of the areas). To avoid the effect of the sex variable, the first name was not reported. All responses were typed to eliminate any possible influence of handwriting (a sample of three of the essay responses is presented in Appendix II).

Construction of Order Factor

The 48 essay responses were then divided into four sets, each set containing 12 essays which were balanced with respect to different levels of each of the three factors (mechanical accuracy, content accuracy, and nationality). Each rater had all the possible combinations of factors levels. The order of essay presentation was kept the same on the basis of their topic, but the other three factors changed from one package to the next. For example, the first essay response began "my sister smokes two packs of cigarettes every day . . . ," which is a partially correct content response; this essay, for example, was presented with many mechanical errors and an American name in the first set. In the second set it was presented with few mechanical errors and a foreign name; in the third set it was presented with many

mechanical errors and a foreign name; and in the last set it was presented with few mechanical errors and an American name. Thus, each of the sets had a different order of essay presentation in terms of combination of levels of three factors (nationality, mechanical accuracy, and content correctness). Table 1 presents the orientation of the order factor across the four sets.

Four dummy papers were added to each package, all with American names, to conceal the balance of the experimental factors; thus each rater only had to score 16 essay responses.

Table 1
Presentation of Order
Across the Four Levels

Essay Presented	O ₁		O ₂		O ₃		O ₄	
	N*	M**	N	M	N	M	N	M
1	F	2	F	1	A	2	A	1
2	F	2	F	1	A	2	A	1
3	F	2	F	1	A	2	A	1
4	A	1	F	2	F	1	A	2
5	A	1	F	2	F	1	A	2
6	A	1	F	2	F	1	A	2
7	A	2	A	1	F	2	F	1
8	A	2	A	1	F	2	F	1
9	A	2	A	1	F	2	F	1
10	F	1	A	2	A	1	F	2
11	F	1	A	2	A	1	F	2
12	F	1	A	2	A	1	F	2

*N = nationality: A = American; F = Foreigner.

**M = mechanical accuracy; 1 = many mechanical errors;
2 = few mechanical errors.

Pilot Studies

Through the process of two pilot studies, using students in educational measurement as raters, necessary changes were made to (1) adjust the content accuracy of some papers closer to the intended level; (2) make sharper distinctions between papers with low mechanical accuracy and high mechanical accuracy; (3) develop scoring instructions to make the raters notice the name of the students before scoring the essay (sample in Appendix III); and (4) develop a scoring guide to be used in the study.

Scoring Procedure

A scoring guide was developed which was thought to deal with important parts of the essay exam being used (sample in Appendix IV). The raters were asked to grade the papers on a nine scale point. The scale given for use was in letter grade, assignment of A to E. A nine point scale for this study was to permit: (1) fine discrimination between papers with different qualities, (2) large possible variability in ratings, and (3) close similarity to the classroom grading scale being used in most universities.

Subjects

For the present study the selected national population was the membership of AERA in Division C (learning and instruction). Initially 300 members drawn by random

sampling, were invited to participate in the study. The study was described as research on scoring essays in educational psychology (a copy of the letter of invitation is presented in Appendix V). A copy of the essay question used in the study was enclosed with a post card for them to return indicating if they agreed to participate in the study.

Forty-four of the initial requests sent out were returned due to incorrect or insufficient addresses. Seventy-seven of the 256 persons contacted agreed to participate in the study. Three of those did not receive the package of essays. Seventy-four packages were sent to those who agreed to participate in this study. Each rater was randomly assigned to receive one of four essay sets (containing 16 essays per set). The subjects were asked to return the scoring sheets in an enclosed stamped envelope. Fifty-nine of seventy-seven subjects who received essay packages returned their answer sheets. For the analysis of this study, fifteen rater responses were needed for each of the four essay sets. The fifty-nine returned packages were not equally distributed across the four order conditions and for two of the order conditions fewer than fifteen ratings were returned; thus for purposes of analysis four additional raters were selected from the faculty in educational psychology of the University of Florida to supplement the data for this research.

Assumptions About the Rating Procedure and Raters

Some assumptions were made about the background of raters and the actual procedure of grading the essays.

It was assumed that

- 1) Raters would have adequate knowledge about the subject matter used in the essay examination.
- 2) Raters would have some experience in grading essay examinations.
- 3) Raters would read the instructions and use the scoring guide given to them.
- 4) Raters would read the name of the essay's writer before assigning a grade to it (as instructed in the instructions).
- 5) Raters would intend to rate only on the basis of accuracy of content.
- 6) Raters would read all the essay responses in the same session.

Design of the Study

The data of this study were analyzed by the split plot factorial design with repeated measures.

Split Plot Design

Order, nationality, content, and mechanics are considered as fixed effects in this study. The rater is the only random effect. Order was the variable used to make

the plots (4 plots), and the raters were nested within the order (15 raters in each plot): i.e., each rater was assigned to only one order.

Rater within order was treated as an independent variable crossed with nationality, content correctness, and the mechanics. This means that each rater received papers with all possible combinations of all factors. Nationality, mechanics, and content were crossed, and this permitted estimation of each main effect and all interactions involving those effects. The twelve repeated measures, obtained from each rater, were the scores assigned to essay responses under the combinations of writer's nationality (2 levels), degrees of mechanical accuracy (2 levels), and degree of content correctness (3 levels). A schematic presentation of the design used in this analysis is presented in Table 2.

Use of repeated measures in a split plot research design has some limitations: According to Kirk (1968), "when matched subjects are assigned to within-block treatment levels, it may be assumed that estimates of treatment effects that have been obtained from the cells are correlated. The model underlying the split plot design will permit a particular kind of statistical dependency between observations in the levels of blocks but requires that error portion of these scores must be independent of each other, and of the treatment effects. There is ample reason to believe that in repeated measures experiments the error components of the

scores are not independent and the variance-covariance matrix departs from the required form" (p. 247). The use of Greenhouse Geisser Conservative F Test was used in case the assumption of symmetry of variance-covariance matrix was violated.

The model used for this analysis is a mixed model:

$X_{NMCR:O}$ = a rating given by a rater receiving the essays in a particular order for nationality (N), mechanics (M), and content (C).

The structural model for this analysis is

$$\begin{aligned} X_{NMCR:O} = & \mu + \alpha_O + \beta_{R:O} + \pi_N + \gamma_C + \theta_M + \alpha\pi_{ON} + \alpha\gamma_{OC} \\ & + \alpha\theta_{OM} + \pi\beta_{NR:O} + \gamma\beta_{CR:O} + \theta\beta_{MR:O} + \pi\gamma_{NC} \\ & + \pi\theta_{NM} + \gamma\theta_{CM} + \pi\gamma\beta_{NCR:O} + \pi\theta\beta_{NMR:O} \\ & + \gamma\theta\beta_{CMR:O} + \alpha\pi\gamma_{ONC} + \alpha\pi\theta_{ONM} + \alpha\gamma\theta_{OCM} \\ & + \pi\gamma\theta_{NCM} + \alpha\pi\gamma_{ONCM} + \pi\theta\gamma_{NMCR:O} \end{aligned}$$

where μ	= the grand mean
α_O	= the effect of O (O=1, . . . n_O)
$\beta_{R:O}$	= the effect of rater nested within order (R=1, . . . n_R)
π_N	= the effect of nationality (N=1, . . . n_N)
γ_C	= the effect of content accuracy (C=1, . . . n_C)
θ_M	= the effect of mechanics (M=1, . . . n_M)

$\alpha\pi$	= the order-by-nationality interaction effect
$\alpha\gamma_{OC}$	= the order-by-content interaction effect
$\pi\beta_{NR:O}$	= the nationality-by-rater interaction effect
$\gamma\beta_{CR:O}$	= the content-by-rater interaction effect
$\theta\beta_{MR:O}$	= the mechanics-by-rater interaction effect
$\pi\gamma_{NC}$	= the nationality-by-content interaction effect
$\pi\theta_{NM}$	= the nationality-by-mechanics interaction effect
$\gamma\theta_{CM}$	= the content-by-mechanics interaction effect
$\alpha\theta_{OM}$	= the order-by-mechanics effect
$\pi\gamma\beta_{NCR:O}$	= the nationality-by-content-by-rater interaction effect
$\pi\theta\beta_{NMR:O}$	= the nationality-by-mechanics-by-rater interaction effect
$\gamma\theta\beta_{CMR:O}$	= the content-by-mechanics-by-rater interaction effect
$\alpha\pi\gamma_{ONC}$	= the order-by-nationality-by-content interaction effect
$\alpha\pi\theta_{ONM}$	= the order-by-nationality-by-mechanics interaction effect
$\alpha\gamma\theta_{OCM}$	= the order-by-content-by-mechanics interaction effect
$\pi\gamma\theta_{NCM}$	= the nationality-by-mechanics-by-content interaction effect

$\alpha\pi\gamma\theta_{\text{ONCM}}$ = the order-by-nationality-by-content-
by-mechanics interaction effect

$\pi\theta\gamma\beta_{\text{NMCR:O}}$ = the rater-by-nationality-by-content-
by-mechanics interaction effect

Summary

For the purpose of the present study an essay question in educational psychology was chosen which dealt with modification of behavior. Twelve essay responses answering the question were written at three different levels of content correctness (highly correct, partially correct, incorrect). Then the essay responses were modified in their mechanical accuracy. Each essay was modified to have many mechanical errors and an exact duplicate of it had few mechanical errors (total of 24 essays). Next, the nationality factor was introduced by coding each essay with a name of an American writer and a duplicate of it with a name of a foreign writer (total of 48 themes). Forty-eight essay responses to this question were composed which carried all possible combinations of nationality (2 levels), mechanical accuracy (2 levels), and content correctness (3 levels). The order factor was introduced by dividing the 48 themes into four sets, each one having a particular order of the combination of factors' levels presented. These were rated by 60 raters, all members of AERA section C (learning and instruction). All raters had all possible combinations of these factors in the essays which they rated. Data from a split plot design with repeated measures were analyzed to investigate the effect of the three factors and their interactions on raters' judgments of essay responses.

CHAPTER IV

RESULTS

Introduction

This study was designed to investigate how some specific paper and examinee characteristics, unrelated to content accuracy, may interact to affect ratings given to essay item responses. The variables of interest in this study were the nationality of examinee, level of mechanical error in the essay response, the order combination, and content correctness. Fifteen research hypotheses were generated to test the possible effects of these variables.

1. There is no significant effect of nationality on raters' judgments of essay item responses.
2. There is no significant effect of mechanical accuracy on raters' judgments of essay item responses.
3. There is no significant effect of content accuracy on raters' judgments of essay item responses.
4. There is no significant effect of order presentation of essays on raters' judgments of essay item responses.
5. There is no significant effect of nationality-by-mechanics on raters' judgments of essay item responses.

6. There is no significant effect of nationality-by-content correctness on raters' judgments of essay item responses.
7. There is no significant effect of mechanics-by-content correctness on raters' judgments of essay item responses.
8. There is no significant effect of mechanics-by-order on raters' judgments of essay item responses.
9. There is no significant effect of content accuracy-by-order on raters' judgments of essay item responses.
10. There is no significant effect of nationality-by-order on raters' judgments of essay item responses.
11. There is no significant effect of nationality-by-mechanics-by-content accuracy on raters' judgments of essay item responses.
12. There is no significant effect of nationality-by-mechanics-by-order on raters' judgments of essay item responses.
13. There is no significant effect of mechanics-by-content accuracy-by-order on raters' judgments of essay item responses.
14. There is no significant effect of nationality-by-content accuracy-by-order on raters' judgments of essay responses.

15. There is no significant effect of nationality-by-content accuracy-by-order-by-mechanical accuracy on raters' judgments of essay item responses.

Results of Four-Way Factorial Analysis from a Split Plot Design

Forty-eight essay responses containing all possible combinations of factors' levels were rated by sixty raters. The data from a split plot design with repeated measures were analyzed by using computer program P2V in BMDP.

Some descriptive data analysis was also performed. The table of cell means and standard deviations for all treatment combinations are reported in Table 3.

The results of the analysis for the split plot factorial design with repeated measures are reported in Table 4. The four-way interaction (NMCO) was significant at a level .05.

Following the recommendations of Kirk (1968), whenever a higher order interaction is significant, the interpretation of lower order interaction(s) and the main effect(s) should be followed very cautiously, and the experiment should be followed by tests of simple main effect(s). The detection of a significant four-way interaction was followed by conducting three separate three-way factorial analyses, one for each level of content correctness. Independent variables in these split plot designs were examinees' nationality and mechanical accuracy. The plotting variable

Table 3

Cell Means and Standard Deviations for the
Four-way Factorial, Split Plot Design

	O_1^*	O_2	O_3	O_4
C_1^{++}	N_1^{**} 7.46 (1.06)	6.46 (2.09)	7.33 (1.08)	7.06 (1.22)
	M_1^+			
	N_2 7.53 (1.59)	7.33 (1.44)	7.20 (1.37)	5.80 (1.89)
	M_2			
	N_1 6.86 (1.76)	7.53 (1.18)	7.26 (1.57)	7.86 (1.24)
	N_2 8.00 (1.30)	7.73 (1.09)	6.66 (1.95)	7.46 (1.35)
C_2	N_1 6.93 (1.79)	3.53 (1.30)	5.20 (1.42)	3.33 (1.17)
	M_1			
	N_2 5.40 (1.40)	3.53 (1.24)	6.33 (2.02)	4.26 (1.94)
	M_2			
	N_1 4.26 (1.79)	4.86 (1.59)	4.13 (2.35)	6.60 (1.68)
	N_2 4.53 (0.91)	7.00 (1.06)	3.93 (1.75)	4.66 (1.12)
C_3	N_1 3.93 (1.16)	3.13 (1.68)	2.46 (1.45)	2.66 (1.23)
	M_1			
	N_2 2.40 (1.24)	2.80 (1.42)	3.53 (1.55)	3.20 (1.82)
	M_2			
	N_1 3.46 (1.35)	2.33 (0.81)	2.73 (1.43)	4.20 (1.93)
	N_2 3.33 (1.17)	4.46 (1.18)	2.66 (1.17)	2.53 (1.18)

* $O_1 - O_4$ = different conditions of essays' order.

**Nationality (N): N_1 = American; N_2 = Foreigner.

†Mechanics (M): M_1 = many errors; M_2 = few errors.

††Content correctness (C): C_1 = highly correct content;
 C_2 = partially correct content; C_3 = incorrect content.

Table 4

Four-way Analysis of Variance from a Split
Plot with Repeated Measures Design

Sources	SS	df	MS	F	Conservative df
N	.138	1	.13	.08	$\theta = \frac{1}{df_N} = \frac{1}{1}$
N•O	54.716	3	18.23	10.81*	
N•R:O(e)	94.477	56	1.68		
M	12.27	1	12.27	6.56*	$\theta = \frac{1}{df_M} = \frac{1}{1}$
M•O	152.60	3	50.86	27.18*	
M•R:O(e)	104.78	56	1.86		
N•M	.35	1	.35	.19	$\theta = \frac{1}{df_{N \cdot M}} = \frac{1}{1}$
N•M•O	74.54	3	24.84	13.45*	
N•M•R:O(e)	103.43	56	1.84		
C	2036.43	2	1018.21	362.20*	$\theta = \frac{1}{df_C} = \frac{1}{2}$
C•O	7.37	6	1.22	.44	
C•R:O(e)	314.85	112	2.81		
N•C	.47	2	.23	.25	$\theta = \frac{1}{df_{N \cdot C}} = \frac{1}{2}$
N•C•O	28.00	6	4.66	4.83*	
N•C•R:O(e)	108.18	112	.96		

Table 4, Continued

Sources	SS	df	MS	F	Conservative df
M·C	1.74	2	.87	.69	$\theta = \frac{1}{\overline{df}_{M \cdot C}} = \frac{1}{2}$
M·C·O	102.17	6	17.02	13.49*	
M·C·R:O(e)	141.41	112	1.26		
N·M·C	.57	2	.28	.26	
N·M·C·O	53.92	6	8.98	8.24*	$\theta = \frac{1}{\overline{df}_{N \cdot M \cdot C}} = \frac{1}{2}$
N·M·C·R:O(e)	122.16	112	1.09		

* Significant at $\alpha = .05$.

**(e) = the error term for preceding effect.

was the treatment combination order. The analyses of data from these designs were done by using GLM procedure in SAS. Before presenting the results of these analyses, and their follow-ups, it is necessary to discuss two statistical adjustments which were required for these data.

Choice of Error Term for Simple Main Effect

It has been noted that the error term for testing the simple main effect in this type of analysis is not always the 'MS_{subject within group}'. The error term changes as different effects are tested (Kirk, 1968). The rule governing the choice of error terms is that if the treatment and interaction, which equal the sum of the simple main effects, have different error terms, the two error terms should be pooled in testing the simple main effect. As an example in a two-way factorial, if the error term for testing main effect A is 'MS_{subject within groups}', and the error term for the AB interaction is 'MS_{bx subject within groups}', then to test the simple main effect, A, it is necessary to pool an error term such as

$$(SS_{\text{sub.w.group}} + SS_{\text{bx.sub.w.groups}}) / (df_{\text{sub.w.group}} + df_{\text{bx.sub.w.groups}})$$

and use the error term in testing the simple main effect of A. For the present study, all error terms necessary for testing simple main effects were calculated and are presented with the analysis results.

Conservative Degrees of Freedom

To test the equality of means of effect in a factorial design requires satisfying an assumption of symmetry of the variance/covariance matrix. Kirk (1968) explains that according to Box (1953), for uncorrelated data, the F test is relatively insensitive to violation of the assumption of homogeneity of variance; however, heterogeneity of both variance and covariances in a design having repeated measures on the same subjects results in a positive bias in the F test. Use of conservative degrees of freedom to test the F statistics has been suggested by Geisser and Greenhouse (1958), Kirk (1968), and Greenwald (1976). Since the design of this study consisted of repeated measures, the conservative degrees of freedom were computed for all the F tests. The conservative test is obtained by multiplying the conventional degrees of freedom by θ which is set equal to $1/(q-1)$. The quantity, $1/(q-1)$, corresponds to the lowest bound that the statistic θ can take (Geisser and Greenhouse, 1958). The $(q-1)$ represents the degrees of freedom associated with the repeated measures effect in the F test. Conservative degrees of freedom associated with tests in four-way analysis of variances are presented along with results of each test. Because the quantity of θ was equal to 1 for the other simple main effect it was not reported in the analysis.

Table 5
Three-way Analysis of Variance for Essays
with Highly Correct Content

Source	SS	df	MS	F
N				
N•R:O + N•C•R:O(e) **	.01	1	.01	.008
	202.65	168	1.20	
M				
M•R:O + M•C•R:O(e)	9.60	1	9.60	6.57*
	246.19	168	1.46	
O				
R:O + R•C:O(e)	6.15	3	2.05	.41
	837.66	168	4.98	
N•M				
N•M•R:O + N•M•C•R:O(e)	.60	1	.60	.44
	225.59	168	1.34	
N•O				
N•R:O + N•C•R:O(e)	22.08	3	7.36	6.13*
	202.65	168	1.20	
M•O				
M•R:O + M•C•R:O(e)	22.70	3	7.56	5.17*
	246.19	168	1.46	
N•M•O				
N•M•R:O + N•M•C•R:O(e)	8.96	3	2.98	2.22
	225.59	168	1.34	

*Significant at $\alpha = .05$.

** (e) = the error term for preceding effect.

Results of Split Plot Design Analysis for
Different Content Accuracy Levels

The analysis of the split plot designs for different content accuracy levels was done by using GLM procedure in SAS.

1. The results of analysis for highly correct content did not show any significant three-way interaction between nationality, mechanics and order. Results of this analysis are presented in Table 5. The two-way interaction between nationality and order was significant at α level .05. The two-way interaction between mechanics and order was also significant at α level .05.

The cell means for nationality-by-order interaction are presented in Table 6.

Table 6

Cell Means for Nationality and Order Combination in
Highly Correct Content Condition

	O_1^*	O_2	O_3	O_4
N_1^{**}	7.16	6.99	7.29	7.46
N_2	7.76	7.53	6.93	6.63

* $O_1 - O_4$ = different conditions of essay order

** Nationality (N): N_1 = American; N_2 = Foreigner.

The follow up of this interaction was to test the effect of nationality at each level of order. This result

is presented in Table 7. The effect of nationality was significant for the fourth order. Thus, when American responses were rated first, the mean rating for Americans (7.46) was significantly greater than the mean for the foreigners (6.63) for highly correct essay responses.

Table 7

Test of Simple Main Effect on Nationality-by-Order Interaction in Highly Correct Content Condition

Source	SS	dF	MS	F
N at O ₁	5.40	1	5.40	4.50
N at O ₂	4.27	1	4.27	3.55
N at O ₃	2.02	1	2.02	1.683
N at O ₄	10.41	1	10.41	8.67*
N·R:O + N·C·R:O _(e) **	202.65	168	1.20	

*Significant at $\alpha = .01$

** (e) = the error term

The cell means for mechanics-by-order interaction are presented in Table 8.

Table 8

Cell Means for Mechanics and Order Combination
in Highly Correct Content Condition

	O ₁ [*]	O ₂	O ₃	O ₄
M ₁ ^{**}	7.49	6.89	7.26	6.43
M ₂	7.43	7.63	6.96	7.66

^{*}O₁ - O₄ = different conditions of essay order.

^{**}Mechanics (M): M₁ = many errors; M₂ = few errors.

The follow up of this interaction was to test the effect of mechanics at each level of order. The results of this analysis are presented in Table 9. The effect of mechanics was significant for the fourth order, where American responses were rated first and the essays with many mechanical errors were rated before essays with few mechanical errors. The mean rating for essay responses with few mechanical errors (7.66) was significantly greater than the mean for essays with few mechanical errors (6.43) for highly correct content essay responses.

2. The results of analysis for partially correct content indicated a significant three-way interaction among nationality, mechanics, and order. The results of this analysis are presented in Table 10. The test of simple main effect was followed by analyzing a split plot with repeated measures design for each level of mechanical accuracy.

Table 9

Test of Simple Main Effect on Mechanics-by-Order
Interaction in Highly Correct Content Condition

Source	SS	dF	MS	F
M at O ₁	.07	1	.07	.047
M at O ₂	8.04	1	8.04	5.50
M at O ₃	1.33	1	1.33	.91
M at O ₄	22.81	1	22.81	15.61*
M·R:O + M·C·R:O(e) **	246.19	168	1.46	

* Significant at $\alpha = .01$

** (e) = the error term.

A. The results of analysis for first level of mechanical accuracy (many mechanical errors) is presented in Table 11. The two-way interaction between nationality and order was significant at α level .05. The cell means for this interaction are presented in Table 12.

The follow up of this significant interaction was to analyze the effect of nationality at each level of order. The result of this analysis is presented in Table 13. The effect of nationality was significant at first and third levels of order. In first level of order where the Americans' essay responses with many mechanical errors were read first, the mean rating for Americans (6.93) was significantly greater than the mean rating for foreigners

Table 10
Three-Way Analysis of Variance for Essays with Partially Correct Content

Source	SS	df	MS	F
N	.60	1	.60	.50
N•R:O + N•C•R:O(e) **	202.65	168	1.20	
M	2.01	1	2.01	1.37
M•R:O + M•C•R:O(e)	246.19	168	1.46	
O	12.48	3	4.16	.83
R:O + R•C:O(e)	837.66	168	4.98	
N•M	.06	1	.06	.04
N•M•R:O + N•M•C•R:O(e)	225.59	168	1.34	
N•O	29.50	3	9.83	8.19*
N•R:O + N•C•R:O(e)	202.65	168	1.20	
M•O	226.68	3	75.56	51.75*
M•R:O + M•C•R:O(e)	246.19	168	1.46	
N•M•O	66.63	3	22.21	16.57*
N•M•R:O + N•M•C•R:O(e)	225.59	168	1.34	

*Significant at $\alpha = .05$.

** (e) = the error term for preceding effect.

Table 11

Two-Way Factorial Analysis of Variance for Essays with Many Mechanical Errors
at Partially Correct Content Condition

Source	SS	df	MS	F
N	.53	1	.53	.41
N•R:O + N•C•R:O + N•M•R:O + N•M•C•R:O(e) **	428.24	335	1.27	
O	162.16	3	54.05	16.78*
M•R:O + R•C:O + M•R:O + M•C•R:O(e)	1083.85	336	3.22	
N•O	33.26	3	11.08	8.72*
N•R:O + N•C•R:O + N•M•R:O + N•M•C•R:O(e)	424.24	336	1.27	

*Significant at $\alpha = .05$.

** (e) = the error term for preceding effect.

Table 12

Cell Means for Nationality and Order Combination for
Essays with Many Mechanical Errors at Partially
Correct Content Condition

	O_1^*	O_2	O_3	O_4
N_1^{**}	6.93	3.53	5.20	3.33
N_2	5.40	3.53	6.33	4.26

* $O_1 - O_4$ = different conditions of essays' order

** Nationality (N): N_1 = American; N_2 = Foreigner.

Table 13

Test of Simple Main Effect on Nationality-by-Order Interaction
for Essays with Many Mechanical Errors at Partially
Correct Content Condition

Source	SS	dF	MS	F
N at O_1	17.63	1	17.63	13.88*
N at O_2	0.00	1	0.00	0.00
N at O_3	9.63	1	9.63	7.58*
N at O_4	6.53	1	6.53	5.14
N·R:O + N·C·R:O + N·M·R:O + N·M·C·R:O (e)	428.24	336	1.27	

* Significant at $\alpha = .01$

** (e) = the error term.

(5.40) for partially correct essay responses. In third level of order where the foreigners' essay responses with many mechanical errors were read first, the mean rating for foreigners (4.26) was significantly greater than the mean rating for Americans (3.33) for partially correct essay responses.

B. The results of analysis for the second level of mechanical accuracy (few mechanical errors) are presented in Table 14. The two-way interaction between nationality and order was significant at α level .05. The cell means for this interaction are presented in Table 15.

Table 14

Two-way Factorial Analysis of Variance for
Essays with Few Mechanical Errors at Partially
Correct Content Condition

Source	SS	dF	MS	F
N	.13	1	.13	.10
N•R:O + N•C•R:O + N•M•R:O + N•M•C•R:O(e)**	428.24	336	1.27	
O	77.00	3	25.66	7.96*
M•R:O + M•C•R:O + R:O + R•C:O(e)	1083.85	336	3.22	
N•O	62.86	3	20.95	16.49*
N•R:O + N•C•R:O + N•M•R:O + N•M•C•R:O(e)	428.24	336	1.27	

*Significant at $\alpha = .05$.

** (e) = the error term for preceding effect.

Table 15

Cell Means for Nationality and Order Combination for
Essays with Few Mechanical Errors at Partially
Correct Content Condition

	O ₁ [*]	O ₂	O ₃	O ₄
N ₁ ^{**}	4.26	4.86	4.13	6.60
N ₂	4.53	7.00	3.93	4.66

* O₁ - O₄ = different conditions of essays' order.

**Nationality (N): N₁ = American; N₂ = Foreigner.

The follow up of this interaction was to analyze the effect of nationality at each level of order. The results of this analysis are presented in Table 16. The effect of nationality was significant at second and fourth levels of order. In the second level of order where the foreigners' essay responses with few mechanical errors were read first, the mean rating for foreigners (7.00) was significantly greater than the mean rating for Americans (4.86) for partially correct essay responses. In the fourth level of order where the Americans' essay responses with few mechanical errors were read first, the mean rating for Americans (6.60) was significantly higher than the mean rating for foreigners (4.60) for partially correct essay responses.

Table 16

Test of Simple Main Effect on Nationality-by-Order
Interaction for Essays with Few Mechanical Errors
at Partially Correct Content Condition

Source	SS	dF	MS	F
N at O_1	.53	1	.53	.41
N at O_2	34.13	1	34.13	26.87*
N at O_3	.30	1	.30	.23
N at O_4	28.03	1	28.03	22.07*
N·R:O + N·C·R:O + N·M·R:O + N·M·C·R:O(e)**	428.24	336	1.27	

*Significant at $\alpha = .01$.

** (e) = the error term.

3. The results of the analysis for the incorrect content level indicated a significant three-way interaction among nationality, mechanics, and order. The results of this analysis are presented in Table 17. The test of simple main effect was followed by analyzing a split plot with repeated measured design for each level of mechanical accuracy.

A. The results of analysis for first level of mechanical accuracy (many mechanical errors) are presented in Table 18. The two-way interaction between nationality and order was significant at α level .05. The cell means for this interaction are presented in Table 19.

Table 17

Three-way Analysis of Variance for Essays with
Incorrect Content

Source	SS	df	MS	F
N				
N.R:O + N.C.R:O(e) **	0.00	1	0.00	0.00
	202.65	168	1.20	
M				
M.R:O + M.C.R:O(e)	2.40	1	2.40	1.64
	246.19	168	1.46	
O				
R:O + R.C:O(e)	6.26	3	2.08	.41
	837.66	168	4.98	
N.M				
N.M.R:O + N.M.C.R:O(e)	.26	1	.26	.19
	225.59	168	1.34	
N.O				
N.R:O + N.C.R:O(e)	31.13	3	10.37	8.64*
	202.65	168	1.20	
M.O				
M.R:O + M.C.R:O(e)	5.40	3	1.80	1.23
	246.19	168	1.46	
N.M.O				
N.M.R:O + N.M.C.R:O(e)	52.86	3	17.62	13.14*
	225.59	168	1.34	

* Significant at $\alpha = .05$.

** (e) = the error term for preceding effect.

Table 18
Two-way Factorial Analysis of Variance for Essays with Many Mechanical
Errors at Incorrect Content Condition

Source	SS	df	MS	F
N				
N•R:O + N•C•R:O +	.13	1	.13	.10
N•M•R:O + N•M•C•R:O(e)**	428.24	336	1.27	
O				
M•R:O + M•C•R:O +	.96	3	.32	.09
R:O + R•C:O(e)	1083.85	336	3.22	
N•O				
N•R:O + N•C•R:O +	29.00	3	9.66	7.60*
N•M•R:O + N•M•C•R:O(e)	428.24	336	1.27	

* Significant at $\alpha = .05$.

** (e) = the error term for preceding effect.

Table 19

Cell Means for Nationality and Order Combination for Essays
With Many Mechanical Errors at Incorrect
Content Condition

	O_1^*	O_2	O_3	O_4
N_1^{**}	3.93	3.13	2.46	2.66
N_2	2.40	2.80	3.53	3.20

* $O_1 - O_4$ = different conditions of essays' order.

** Nationality (N): N_1 = American; N_2 = Foreigner.

The follow up of this significant interaction was to analyze the effect of nationality at each level of order. The results of this analysis are presented in Table 20. The effect of nationality was significant at first and third levels of order. In first level of order where the Americans' essay responses with many mechanical errors were read first, the mean rating (3.93) was significantly greater than the mean rating for foreigners (2.40) for incorrect content essay responses. In third level of order where the foreigners' essay responses with many mechanical errors were read first, the mean rating for foreigners (3.53) was significantly greater than the mean rating for Americans (2.46) for incorrect content essay responses.

Table 20

Test of Simple Main Effect on Nationality-by-Order Interaction
for Essays with Many Mechanical Errors at
Correct Content Condition

Source	SS	dF	MS	F
N at O ₁	17.63	1	17.63	13.88*
N at O ₂	.83	1	.83	.65
N at O ₃	8.53	1	8.53	6.71*
N at O ₄	2.13	1	2.13	1.67
N•R:O + N•C•R:O + N•M•R:O + N•M•C•R:O (e)	428.24	336	1.27	

* Significant at $\alpha = .01$

** (e) = the error term.

B. The results of the analysis for the second level of mechanical accuracy (few mechanical errors) are presented in Table 21. The two-way interaction between nationality and order was significant at α level .05. The cell means for this interaction are presented in Table 22.

The follow up of this interaction was to test the effect of nationality at each level of order. The results of this analysis are presented in Table 23. The effect of nationality was significant at second and fourth levels of order. In second level of order where the foreigners' essay response with few mechanical errors were read first,

Table 21
Two-way Factorial Analysis of Variance for Essays
with Few Mechanical Errors at Incorrect Content Condition

Source	SS	df	MS	F
N	.13	1	.13	.10
N•R:O + N•C•R:O + N•M•R:O + N•M•C•R:O(e)**	428.24	336	1.27	
O	10.70	3	3.56	1.10
R:O + R:O + M•R:O + M•C•R:O(e)	1083.85	336	3.22	
N•O	55.00	3	18.33	14.43*
N•R:O + N•C•R:O + N•M•R:O + N•M•C•R:O(e)	428.24	336	1.27	

*Significant at $\alpha = .05$.

** (e) = the error term for preceding effect.

Table 22

Cell Means for Nationality and Order Combination
for Essays with Few Mechanical Errors at Incorrect
Content Condition

	O_1^*	O_2	O_3	O_4
N_1^{**}	3.46	2.33	2.73	4.20
N_2	3.33	4.46	2.66	2.53

* $O_1 - O_4$ = different condition of essays' order.

** Nationality (N): N_1 = American; N_2 = Foreigner.

Table 23

Test of Simple Main Effect on Nationality-by-Order
Interaction for Essays with few Mechanical Errors
at Incorrect Content Condition

Source	SS	dF	MS	F
N at O_1	.13	1	.13	.10
N at O_2	34.13	1	34.13	26.87*
N at O_3	.03	1	.03	.02
N at O_4	20.83	1	20.83	16.40*
N•R:O + N•C•R:O				
N•M•R:O + N•M•C•R:O (e)**	428.24	336	1.27	

* Significant at $\alpha = .01$.

** (e) = the error term.

the mean rating for foreigners (4.46) was significantly greater than the mean rating for Americans (2.33) for incorrect content essay responses. In fourth level of order where the Americans' essay responses with few mechanical errors were read first, the mean rating for Americans (4.20) was significantly greater than the mean rating for foreigners (2.53) for incorrect content essay responses.

Summary

Sixty raters graded 48 essay responses. These 48 essay responses were divided into four sets, 12 essays in each set; each rater rated 12 essays which had all possible combinations of four variables' levels. These variables were examinees' nationality, order of essays read, content accuracy, and mechanical accuracy. Data from a 4-way factorial split plot with repeated measures design were analyzed. The four-way interaction among the nationality, order, content, and mechanics was significant. The follow-up of this four-way significant interaction was to test the effect of mechanics, order presentation, and nationality at each level of content correctness. When the essays were highly correct the effect of mechanics was significant at only one level of order, and the effect of nationality was significant at four levels of order. When the essays were partially correct and had many mechanical errors, the effect of nationality was significant at first and third levels of order. When the essays were partially correct and had

few mechanical errors, the effect of nationality was significant at second and fourth levels of order. When the essays were incorrect and had many mechanical errors the effect of nationality was significant at first and third levels of order. When the essays were incorrect and had few mechanical errors, the effect of nationality was significant at second and fourth levels of order.

CHAPTER V

DISCUSSION

The purpose of this study was to investigate how some variables, unrelated to content accuracy, interact in raters' judgments of essay responses. The variables of interest were examinees' nationality, essay responses' mechanical accuracy, and the content correctness of essay responses. The effect of these variables was tested for different presentation orders of the three treatment combinations. Forty-eight essay responses covering all possible treatment combinations were rated by sixty raters.

The four-way interaction among order, nationality, content and mechanics was significant. Since the primary interest of this study was not to examine the differences between the three levels of content correctness, and it was assumed that the raters will rate essays at different content correctness levels differently, the follow up of this interaction was to test the effect of the order, mechanics and nationality at each level of content correctness. The discussion of results will be divided into two sections: discussion of mechanics effect and discussion of nationality effect.

Mechanics

When the essays responses were highly correct in content, there was a significant interactive effect of mechanics and order. Tests of this interaction revealed that mechanics were significantly affecting the raters' judgments when they were read in fourth order of presentation. The fourth order of presentation was that all American examinees' essay responses with many mechanical errors were presented first, then followed by all American examinees' essay responses with few mechanical errors, followed by all foreign examinees' essay responses with many mechanical errors and last with foreign examinees' essay responses with few mechanical errors were presented. In this situation the higher mean was assigned to essays with few mechanical errors. This finding is in agreement with findings of Scannel and Marshall (1966), Freedman (1979), and Chase (1979) who have found that factors such as compositional errors, grammatical errors, and weak sentence structure will affect the raters in their judgments of essay responses.

When the essay responses were partially correct or incorrect in content, the interactive effect of mechanics, nationality, and order was significant. These interactions were tested by examining the effect of nationality-by-order interaction at each level of mechanics; for this reason the mechanics effect was not testable at the partially correct and incorrect content levels.

Nationality

The effect of nationality was significant at all three levels of content correctness for different levels of mechanics and order combinations. Figure 1 presents all significant nationality effects at different levels of content correctness, mechanics, and order. The notation "A vs. F" indicates the mean for responses coded with American names was significantly greater than the mean for responses coded with foreign names. The notation "F vs. A" indicates the mean was greater for responses coded with foreign names was significantly greater than means for responses coded with American names.

Examination of the first significant nationality effect from Figure 1 indicates that higher mean ratings were assigned to American examinees when the essays read were presented in fourth type of order. In this particular order presentation of essays all American examinees' essays are presented in the first half of the set and are before the foreign examinees' essays.

The second significant nationality effect occurred when the essays were partially correct and had many mechanical errors. This significant effect occurred when essays were read in first and third levels of order presentation. In first order presentation the higher mean ratings were assigned to American examinees; in this condition the American examinees' essay with these characteristics (partially correct content and many mechanical errors)

were presented before foreign examinees' essays with the same characteristics. In third order presentation condition the higher mean ratings were assigned to foreign examinees' essays; in this condition the foreign examinees' essays with these characteristics (partially correct content and many mechanical errors) were presented before American examinees' essays with the same characteristics.

The third significant nationality effect occurred when the essays were partially correct and had few mechanical errors. This significant effect occurred when essays were read in second and fourth levels of order presentation. In second order of presentation the higher mean was assigned to foreign examinees' essays; in this condition the foreign examinees' essay with these characteristics (partially correct and few mechanical errors) is presented before American examinees' essay with the same characteristics. In fourth order presentation condition the higher mean ratings were assigned to American examinees' essays; in this condition the American examinees' essays with these characteristics (partially correct content and few mechanical errors) are presented before foreign examinees' essays with the same characteristics.

The fourth significant nationality effect occurred when the essays were incorrect in content and had many mechanical errors. This significant effect occurred when essays were read in first and third levels of order presentation. In the first order presentation the higher mean

rating was assigned to American examinees; in this condition the American examinees' essays with these characteristics (incorrect content and many mechanical errors) were presented before foreign examinees' essays with the same characteristics. In third order condition the higher mean rating was assigned to foreign examinees' essays; in this condition the foreign examinees' essay with these characteristics (incorrect content and many mechanical errors) was presented before American examinees' essay with the same characteristics.

The fifth significant nationality effect occurred when the essays were incorrect in content and had few mechanical errors. This significant effect occurred when essays were read in second and fourth levels of order presentation conditions. In the second order the higher mean rating was assigned to foreign examinees' essays; in this condition the foreign examinees' responses with these characteristics (incorrect content and few mechanical errors) were presented before American examinees' essays with the same characteristics. In the fourth order, the higher mean rating was assigned to American examinees' essays; in this condition the American examinees' responses with these characteristics (incorrect content and few mechanical errors) were presented before foreign examinees' essays with the same characteristics.

	O ₁	O ₂	O ₃	O ₄
	A vs. F	F vs. A	A vs. F	F vs. A
	F vs. A	A vs. F	F vs. A	A vs. F
1-highly correct content				*
2-partially correct content & many mech. errors	*		*	
3-partially correct content and few mech. errors		*		*
4-incorrect content & many mech. errors	*		*	
5-incorrect content & few mech. errors		*		*

Figure 1

Significant Nationality Effects

A vs. F = the mean for responses coded with American names was significantly greater than the mean for responses coded with Foreign names.

F vs. A = the mean for responses coded with Foreign names was significantly greater than the mean for responses coded with American names.

Conclusion

The present study showed that when the essays are highly correct in content, the rating judgment was affected by mechanics of writing. This finding occurred only when essays were presented in a particular order; all papers written by American examinees were presented following each other and before the papers written by foreign examinees.

The examinees' nationality affected these raters judgments under five conditions; however, when the direction of the mean difference was evaluated it was revealed that these effects were mainly caused by the order in which the essays were read: in every one of these five conditions the higher mean rating was assigned to the nationality presented first in the set of essays read. Thus, even though the present study has shown a significant nationality effect, this effect was mainly caused by presentation order of the essays; with a different presentation order the effect of nationality might be different. The effect of order in which the essays are read has been noted by McColly (1970) as a possible source of error in essay grading. Hales and Tokar (1975) and Hughes et al. (1980) have studied the context effect and have found that context in which the essays are read affects the rating judgment of the essays. The final conclusion for the present study is that in situations which require grading essay examinations written

by examinees of different nationality, the order in which the essays are graded is a source of error and should be considered seriously.

Classroom teachers with mixed nationalities in their classroom should consider the conditions under which the foreign students are seated in the classroom. It is possible that foreign students may sit in a group and this will determine the position of their essays in a set of responses collected from the class. Teachers should also consider that due to foreigners' difficulties in writing, these students may tend to return their responses after other students have submitted their papers and this will also determine the order of the foreign students' responses in the set of responses to be graded.

Some practical suggestions to reduce the effect of order on essay response scores in classrooms with mixed nationalities would be

1. mix the essay responses after collection;
2. judge the essay responses at least twice, once starting from the first essay to the last and again in reverse order;
3. If there is more than one question, grade all the responses to a question at a time, and randomly reorder the papers before grading the next question.

Limitations of the Study

There are two limitations of this study which should be considered before the results are generalized:

1. One limitation of this study is imposed by sample. The raters chosen for this study were professional educational psychologists who were asked to grade essays written to answer an educational psychology question. It seems reasonable to believe that the raters have had some educational background in measurement and may be more aware of factors that affect the reliability of grades given to essay responses. It also seems reasonable to believe that the factors unrelated to content may not affect their judgments of essay responses to the same degree that it would affect raters in other disciplines. Thus the results of this study may not be generalizable to instructors and essay responses in other disciplines.
2. Another limitation of this study is that the raters were not grading essay responses of examinees whom they knew. The examinees, their past achievement and experiences were unknown to the raters. In many essay grading situations the raters know the examinees, and have some expectations of students' achievements and abilities. As a result of this

limitation of findings of this study may not be generalizable to classroom situations where raters have close personal contact with students.

Suggestions for Future Research

The results of this experimental study have revealed that rating essay responses objectively is very difficult. The variables examinees' nationality, mechanics of essay response, content correctness of response, and context of essay responses were studied. These variables generally had interactive effects on raters' judgments. One other variable which might have an interactive effect on raters' judgments of essay responses is the examinee's handwriting. Handwritten essay responses will be closer to the actual situation which a classroom will have in grading essay responses.

Raters' experience has been mentioned by some researchers to be a significant factor in rating of essay responses. For future study in this area it is recommended to consider the raters' experience as a factor which might change the findings. It would be best if the raters have common experience and are given training in essay scoring before the experiment is conducted.

The measurements in the present study were based on quantitative impressions which the raters assigned to essay responses; their qualitative impressions of the responses and how they arrived at the grades assigned to

essay responses were not considered. The collection of some qualitative reactions from the raters, by interview or a questionnaire, after they have rated the essay responses might help to reveal if their possible biases are caused by the rater's belief or they are caused by some unconscious influence from unrelated variables in the essay response.

The essay responses used for the present study were fictitiously generated. For future study it would be interesting to use real students' responses. This will create a more varied writing pattern and might result in a sharper distinction between the writing pattern of a native English speaker and the writing pattern of a foreigner.

Summary

The effect of mechanics was significant when the essays were highly correct in content and essays were read in a special order. In this order all the American examinees' responses were read before the foreign examinees' responses, and essays with many mechanical errors were read before the essays with few mechanical errors.

The effect of nationality was significant under several conditions; however, the higher mean was always assigned to the nationality which appeared first in the order of presentation.

An implication of these findings for scoring essay examinations in higher education is in a situation requiring grading examinations written by examinees of different nationalities. The order in which the essays are graded is a source of error and should be considered seriously.

Possible future research topics suggested by these findings include effects of handwriting, rater's experience, use of qualitative impressions of raters, and use of actual responses of examinees of different nationalities.

APPENDIX I

SAMPLE OF ESSAY ITEM USED IN THE STUDY

ESSAY QUESTION

Suppose your roommate is a smoker. You would like to change the frequency of cigarette smoking. Identify specific, step-by-step procedures you would follow to obtain the desired change by using shaping strategy. Use technical terms where they are appropriate.

APPENDIX II

SAMPLE OF SOME ESSAY ITEM
RESPONSES USED IN THE STUDY

SAMPLE OF HIGHLY CORRECT CONTENT ESSAY WITH FEW
MECHANICAL ERRORS CODED WITH FOREIGNER NAME

Name KASRAVI, S
S.S. 000 03 3052
Date MAY 1, 1980

My roommate now smokes 2 packs of cigarettes a day. He has been smoking for a long time. I would like to decrease the number of cigarettes that he smokes. We are also saving money to buy a stereo for our apartment.

I know that he wants the stereo very badly and he can't get it without my help. I can use saving money as a reinforcer. I know that he can reduce the number of cigarettes to 10 per day (objective).

I will discuss the plan with him and will let him know that he has to reduce the number of cigarettes that he smokes in a day to 30/day in the first week and to 20-25 cigarettes/day in the second week, and to 15 cigarettes in the third week and to 10 cigarettes in the fourth week.

I will tell him that he gets the reinforcer at the end of each day if and only if he has attained the daily goal. And I will also explain that the reinforcer is that we each put one dollar in our savings for the stereo. At the end of each week if the plan was perfectly implemented we would put one more dollar into the saving account, if not we make the needed changes.

SAMPLE OF INCORRECT CONTENT ESSAY WITH MANY
MECHANICAL ERRORS CODED WITH AMERICAN NAME

Name JONES, L

S.S. 155 65 1544

Date MAY 1, 1980

My roommate smokes and all I want to do is to decrease the number of cigaretts that he smokes in a day.

I tell him I will gave him a querter each time that a day passes and he has not smoked more than 20 cig. The week after that I tell him I will give him a quarter if he does not smoke more than 10 cig. a day.

This should work. Because all I want to do is shape his smoking so that I will decrease the number of cigaretts he smokes a day.

SAMPLE OF PARTIALLY CORRECT CONTENT ESSAY WITH MANY
MECHANICAL ERRORS CODED WITH AMERICAN NAME

Name FORD,N

S.S. 138 03 9739

Date MAY 1, 1980

My sister smokes 15 cigarettes a day (baseline) I know that she likes to go to the movies So I tell her that if she cuts down on the number of cigarettes that she smoks in a week, like to 15/day for the first week and 10/day the second week and 5/day in the third week, we will go the movies at the end of each week (approximation plan). I would measured her accomplishment on a weekly basis, at the end of week if she had not had more than the amount we aggreed on, we would go to a movies (reinforcment). I am not sure that going to a movie is a strong enough reinforcement or even if the plan is correct or not, but at the end of each week I will evaluate the progress and make the aprobate changes.

APPENDIX III

SAMPLE OF SCORING INSTRUCTION

INSTRUCTIONS

1. Read the class description, essay question and scoring guide.
2. Go through the theme pocket and check that the names on the student responses match those on your grading sheet. If any mismatches or duplicates occur, please note them on the grading sheet.
3. Turn to the first essay response, locate the student's name on your grading sheet.
4. After reading the students' response, record your holistic rating on the grading sheet next to the student's name and ID#.
5. Continue this process until all essays are scored.
6. After all themes are scored, place only the grading sheet in the enclosed stamped envelope and return to us.

THANK YOU

CLASS DESCRIPTION

The enclosed responses are randomly selected from the total population of students' responses in an introduction to developmental psychology class. Total population of class was composed of 80% white, 20% black, 55% female, 45% male. About 15% of students in this class were foreign students.

APPENDIX IV

SAMPLE OF SCORING GUIDE USED FOR RATING
OF ESSAY ITEM RESPONSES IN THE STUDY

CRITERION FOR RATING ESSAY ANSWERS

Dear Rater: these are guidelines on which you can base your judgment for rating the essay exams. Assign an overall grade to each response using nine point grading scale:

A, A-, B, B-, C, C-, D, D-, and E.

An outstanding answer (for grade A) is required to have all the steps below.

1. Collect the baseline: frequency that behavior appears, conditions under which the behavior appears.
2. Select the terminal behavior: the desired changed behavior.
3. Select appropriate reinforcer (practical, available, powerful).
4. Initial plan: list successive approximations of the terminal behavior beginning with the initial behavior (note that they stay with plan).
5. Implementing the plan:
 - A. tell the plan to the person;
 - B. reinforce each occurrence of starting behavior;
 - C. do not move to new approximation until the person has mastered the previous approximation.
6. Evaluation: compare frequency during implementation phase with baseline. Evaluate the reinforcer. If necessary change the plan and/or reinforcer.

APPENDIX V

SAMPLE OF THE LETTER OF INVITATION
FOR PARTICIPATION IN THE STUDY

Dear

In Foundations of Education at the University of Florida we are engaged in a study of factors relating to student achievement in baccalaureate teacher education programs. One course that is universally required in such programs is introductory educational psychology. Traditionally students in such courses have been evaluated on a variety of criteria, including both essay and objective examinations. Thus one phase of our study requires us to gather expert ratings of student responses to a sample of essay items from educational psychology.

As a member of AERA, Division C, you were selected as a qualified expert who might be willing to rate a small number of essay responses from the content area of learning and instruction. Your participation would involve reading a selection of 16 one-page responses to a single essay item and assigning an overall rating to each paper. A copy of the essay item and scoring key are enclosed. In a pilot study, most raters were able to complete this task in less than thirty minutes and found it interesting.

If you would be willing to assist us in this effort, please check and return the enclosed post card. We will then send you a packet of essay responses for grading. Your assistance represents a valuable service to our project and we regret that it is not possible to pay you for your participation; we would, however, be happy to furnish you with summaries of findings as they become available. You may indicate your interest in receiving such summaries on the return post card.

Thank you for considering this request for your professional assistance. We look forward to hearing from you.

Sincerely,

Linda Crocker
Associate Professor and Head,
Research Section
Foundations of Education

LC:efr

REFERENCES

- Barr, A., Goodnight, J., Sall, J., Blair, W., and Chilko, D. SAS user's guide. Raleigh, North Carolina: SAS Institute, Inc., 1979.
- Bean, K.L. Construction of educational and personal tests. New York: McGraw-Hill Book Inc., 1953.
- Box, G.E.P. Non-normality and tests on variances. Biometrika. 1953, 40, 318-335.
- Braddock, R., Lloyd-Jones, R., and Schoer, L. Research in written composition. Champaign, Illinois: National Council of Teachers of English, 1963.
- Clifford, M.M., and Walster, E. Research note: The effect of physical attractiveness on teacher expectations. Sociology of Education. 1973, 46(2), 248-158.
- Chase, C.I. The impact of some obvious variables on essay test scores. Journal of Educational Measurement. 1968, 5(4), 315-319.
- Chase, C.I. Test of English as a foreign language. In Buros, O.K. (ed.), The seventh mental measurements yearbook (Vol. II). Highland Park, New Jersey: The Gryphon Press, 1972.
- Chase, C.I. The impact of achievement expectations and handwriting quality on scoring essay tests. Journal of Educational Measurement. 1979, 6(1), 39-42.
- Diener, T.J., and Kerr, L. Institutional responsibilities to foreign students. The New Direction for Community Colleges. 1979, 26, 49-56.
- Dixon, W.J., and Brown, M.B. (eds.) Biomedical computer program P-series. University of California Press, University of California, 1979.
- Dubois, P.H. A test-dominated society: China, 1115 B.C.-1905 A.D., in Anastasi, A. (ed.). Testing problems in perspective. Princeton, New Jersey: Educational Testing Service, 1965.

- Ebel, R.L. Essentials of educational measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.
- Freedman, Sarah W. How characteristics of student essays influence teachers' evaluations. Journal of Educational Psychology. 1979, 71(3), 328-338.
- Geisser, S., and Greenhouse, S.W. An extension of Box's results on the use of the F Distribution in multivariate analysis. Annals of Mathematical Statistics. 1958, 29, 885-891.
- Gipps, G., and Ewen, E. Scoring written work in English as a second language. British Journal of Educational Psychology. 1974, 44(1), 121-125.
- Greenwald, A.G. Within subject designs: To use or not to use? Psychological Bulletin. 1976, 83(2), 314-320.
- Hales, Lloyd, W., and Tokar, E. The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. Journal of Educational Measurement. 1975, 12(2), 115-117.
- Harris, Winifred H. Teacher response to student writing: A study of the response patterns of high school English teachers to determine the basis of teacher judgment of student writing. Research in the Teaching of English. 1977, 11(2), 175-185.
- Hogan, Thomas P., and Mishler, Carol J. Judging the quality of student writing: When and how. The Elementary School Journal. 1979, 79(3), 142-146.
- Hosley, Deborah, and Meredith, Keith. Inter- and intra-test correlates of the TOEFL, TESOL Quarterly. 1979, 13(2), 209-217.
- Huck, Schuyler, and Bounds, William. Essay grades: An interaction between grader's handwriting clarity and the neatness of examination papers. American Educational Research Journal. 1972, 9(2), 279-283.
- Hughes, D.C., Keeling, Brian, and Tuck, Bryan F. The influence of context position and scoring method on essay scoring. Journal of Educational Measurement. 1980, 17(2), 131-135.
- Kirk, R. Experimental design: Procedures for the behavioral sciences. Belmont, California: Wadsworth, 1968.
- Klein, S.P., and Hart, F.M. Chance and systematic factors affecting essay grades. Journal of Educational Measurement. 1968, 5(3), 197-206.

- McColly, William. What does educational research say about the judging of writing ability? Journal of Educational Research. 1970, 64(4), 148-156.
- Manganel, Richard Arthur. Reading ease as a factor in the grading of high school expository writing (Ed.D. dissertation, State University of New York at Buffalo, 1968). Dissertation Abstracts International, 1968, 29, 836A (University Microfilms No. 68-12454).
- Marshall, J.C. Composition errors and essay examination grades re-examined. American Educational Research Journal. 1967, 4(4), 375-385.
- Marshall, J.C., and Powers, J.M. Writing neatness, composition errors and essay grades. Journal of Educational Measurement. 1969, 6(2), 97-101.
- Payind, M.A. Academic, personal and social problems of Afghan and Iranian students in the United States. Educational Research Quarterly. 1979, 4(2), 3-11.
- Renner, R.R. Students on the hoof. International Education Newsletter. College of Education, University of Florida, May 1981.
- Robertson, Piedad F., and Rogers, J. Teaching English as a second language. New Directions for Community Colleges. 1979, 26, 57-65.
- Scannell, D.P., and Marshall, J.C. The effect of selected composition errors on grades assigned to essay examinations. American Educational Research Journal. 1966, 3(2), 125-130.
- Snibbe, Homa Mahmoudi. The relationship of professors' personality characteristics to grading behaviors (Ph.D. dissertation, University of Utah, 1970). Dissertation Abstracts International, 1970, 31, 3354A. (University Microfilms No. 71-00927).
- Starring, Robert W. A study of ratings of comprehensive examination themes when certain elements are weakened (Ed.D. dissertation, Michigan State College, 1952). Dissertation Abstracts International, 1952, 12, 850. (University Microfilms No. 00-04327).
- Theimer, W.C. The relationship between certain personality and educational factors and the grading pattern of secondary school teachers (Ph.D. dissertation, University of Denver, 1967). Dissertation Abstracts International, 1967, 28, 2922A. (University Microfilms No. 68-02403).

- Troyanovich, John. How defensible is writing as an objective in short-term foreign language experience? Foreign Language Annals. 1974, 7(4), 435-442.
- Ure, Jean. Education in a second language: Planning a remedial service for advanced students. English Language Teaching. 1974, 29(3), 229-239.
- Wen, Shin-Sung. Racial halo on evaluative rating: General or differential? Contemporary Educational Psychology. 1979, 4(1), 15-19.
- Wilkinson, G.P. The three r's: Are they being neglected. Education Canada. 1974, 14(4), 22-30.
- Wong, Martin. Scoring guide for improving essay reliability. Improving College and University Teaching. 1973, 21(1), 26-28.
- Yeasmeen, Nazma, and Baker, Donald G. A half century of research of essay testing. Improving College and University Teaching. 1973, 21(1), 24-25.
- Zimmerman, June Woodard. An investigation of the effect of students' self-concept of writing ability on their writing processes and writing achievement: Case studies of eleven black college freshmen (Ph.D. dissertation, Florida State University, 1977). Dissertation Abstracts International, 1977, 38, 7132A. (University Microfilm No. DDK 78-08979).

BIOGRAPHICAL SKETCH

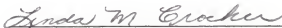
Behrokh Ahmadi was born on April 5, 1953, in Ahwaz, Iran. She graduated from Kharazmi High School on June 1971 in Tehran, Iran. She attended National University of Iran in September 1971, and graduated in January 1975 with a Bachelor of Science degree in psychology.

In January 1976 she began her graduate studies in the College of Education at University of Florida. She received the degree of Master of Education in June 1977.

Behrokh began her doctoral program at the University of Florida in the fall of 1977. While working on her doctoral degree she has held at various times, a graduate teaching assistantship in the area of research design consultation, a graduate assistantship in the area of data management and statistical analysis, and a graduate research assistantship in the area of evaluation coordination.

Behrokh is a member of the American Educational Research Association. She has written a paper and has coauthored a paper which have been presented at annual meetings of the American Educational Research Association.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.


Linda M. Crocker, Chairperson
Associate Professor Foundations
of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

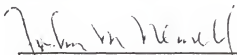

Robert S. Soar
Associate Professor of
Foundations of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.


Walter Cunningham
Associate Professor of
Psychology

This dissertation was submitted to the Graduate Faculty of the Department of Foundations of Education in the College of Education in the College of Education and to the Graduate Council, and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

May 1982



Chairman, Foundations of
Education

Dean for Graduate Studies
and Research